**Using machine learning for automatic censoring**

We propose the use of neural networks for the purpose of automatically censoring explicit content from audio.

We would use different end to end speech recognition models such as sequence to sequence model and also look into using CNNs for the same as they are computationally less expensive. Our aim is to get results comparable to traditional techniques such as LSTMs while improving the speed at which the model runs.

It can be advantageous as it can drastically reduce manual labour.

**Related works:**

**SIFT features on spectrogram images:** We will be using a similar approach as the following paper. It uses SIFT features on spectrograms to make them invariant to transformations.This would allow us to process speech at different paces.We shall use this instead of other methods such as cnns in order to cut down on computation where possible. The model used naive bayes clustering on top of SIFT to perform speech classification but we would require a more complex model for our purposes.

**Conformer: Convolution-augmented Transformer for Speech Recognition :** This paper proposes a new block which combines Convolution layers and Transformer layers to capture both long and short length interactions. These blocks have been shown to provide better accuracy and also are less computationally expensive at smaller sizes than traditional methods.

**On the limit of English conversational speech recognition:** This paper works on attaining best performance in speech recognition.  
It tries different optimizers and models and compares them on the basis of its accuracy and speed. It shows that AdamW optimiser with conformers as encoder and LSTMs as decoders give the best result for their task. Though their task was different than ours, it still acts as a valuable starting point to build upon.

**Basic ingredients:**

We will require a large dataset consisting of audio snippets accompanied by a series of labels.

Each audio clip will be broken down further into smaller slides before using it in the model.Each of these slides will have its label corresponding to whether it contains explicit content or not.

We will need to determine the hyperparameters of our model, namely, the number of attention heads in the conformer blocks, the learning rates of the optimizer, kernel size in the convolution layer of the conformer block to name a few.

**Model:**

SIFT will be applied to the spectrogram of each audio slide and the data points are used as an encoding. We may need to add paddings to them as conformers require a fixed length block. A convolution layer can be applied to downsample in case it is required due to length limits in conformer blocks because its time complexity is quadratic.This will form the encoder model.

For the decoder, we will be using LSTMs to classify each slide as whether it contains explicit content or not. This can then be used to cut audio corresponding to the specified slides.

We will try to make the slides as small as possible without trading it for accuracy to make the cut as close to the words as possible.

**Evaluation:**

To evaluate the model we shall record the number of incorrect classification of slides as a measure to test accuracy. Our aim is to make the model able censor the audio as precisely as possible without cutting anything extra.The size audio slices were reduced to as small as possible without dropping the accuracy below 90%.

Though the model failed to attain real time performance it was faster than the traditional method which gave a real time factor as opposed to what we achieved.